NTP
National Toxicology Program

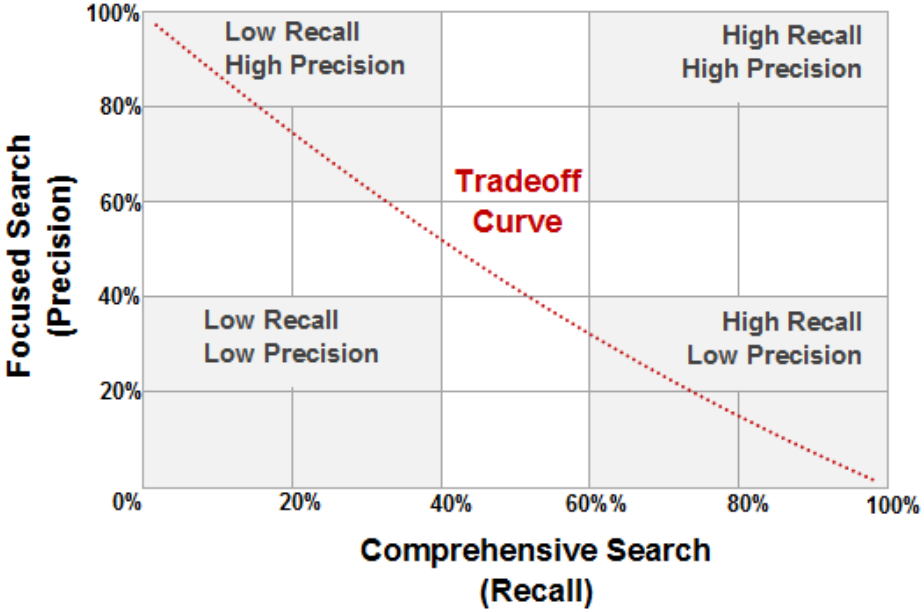
SWIFT: A Text-mining Workbench for Systematic Review

Ruchir Shah, PhD
Sciome LLC

NTP Board of Scientific Counselors Meeting
June 16, 2015

Large Literature Corpus: An Ever Increasing Challenge

- Systematic review begins with a literature search
- To achieve high recall, search needs to be comprehensive
- Comprehensive search = large literature corpus



	PFOS/PFOA	BPA	Transgenerational
PubMed	**6,331**	**7,700**	**48,638**

Large Literature Corpus: An Ever Increasing Challenge

Starting with a large literature corpus, reviewers typically face the following questions:

[A] What are the major topics being discussed in this corpus?

- What Health Outcomes, Exposures, Tox21/EDC chemicals?
- How can I quickly identify scientific areas that are data poor/rich?

Rapid review (Scoping Report)
Problem formulation or Topic refinement

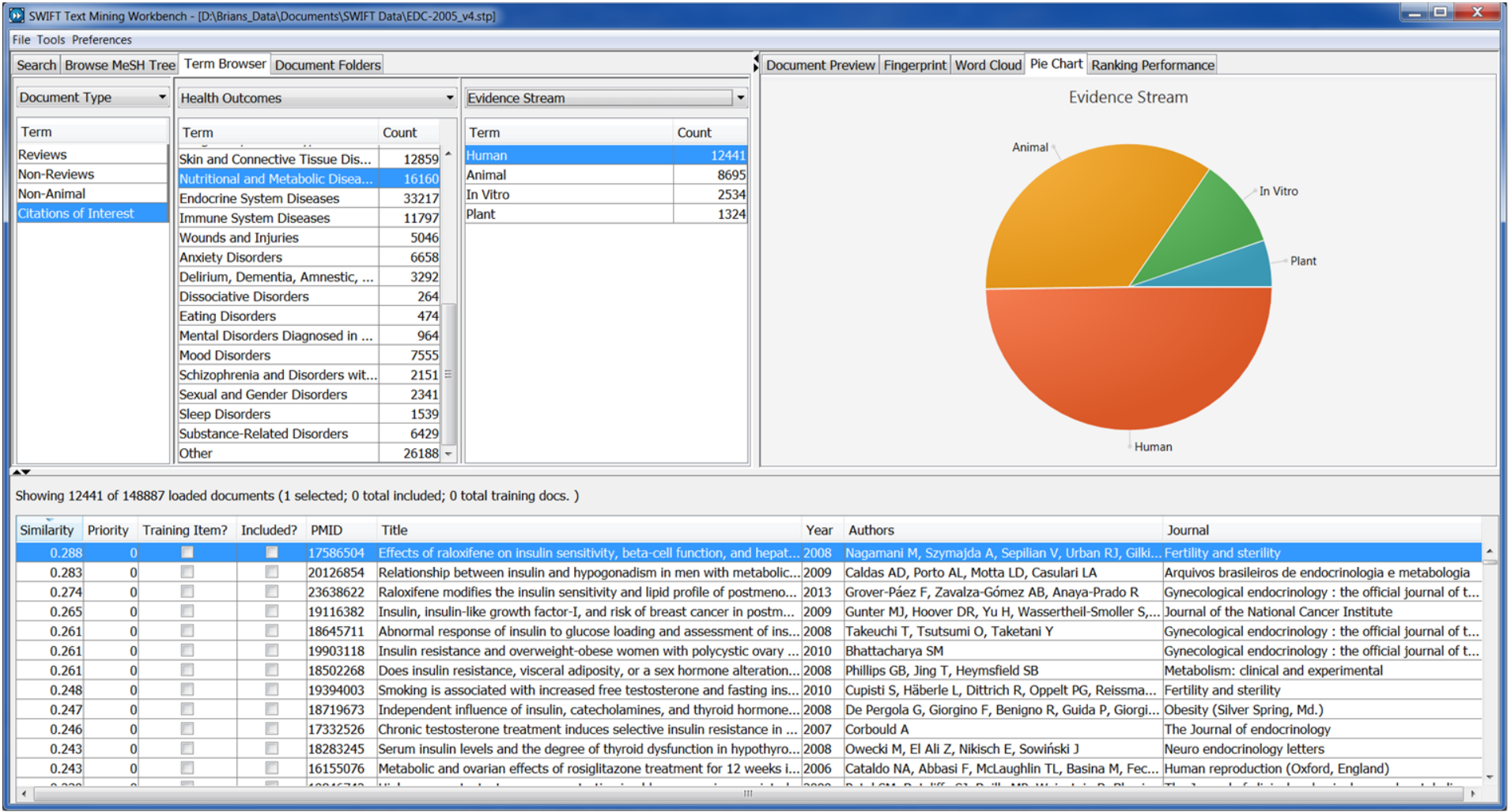
[B] How to facilitate the manual screening to make it efficient?

- Is it possible to prioritize documents prior to screening?
- Would it be possible to reduce screening burden without missing critical information?

Make literature screening more efficient

Outline

- Text-mining, information retrieval and machine learning to statistically analyze large collections of documents
- Provide a user friendly and interactive workbench



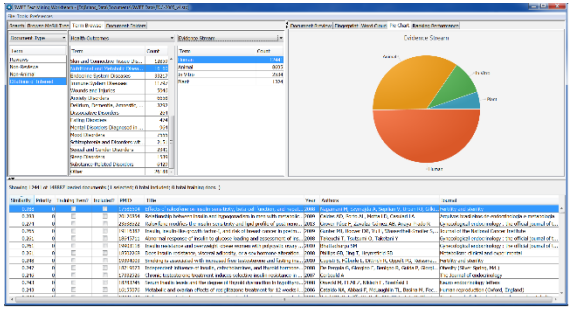
SWIFT

PubMed IDs
from search

Load

Load project
in SWIFT

Analyze



Easily produce:
charts,
statistics,
prioritized list &
more…

1. Organize, Explore and Filter, using “Tags”

“Tagging” means automatically assigning relevant annotations to a document.

- **SWIFT tags documents according to** :
  - Health Outcomes
  - Evidence Stream
  - Exposure
  - Tox21 Chemicals
  - MeSH Headings
  - MeSH Supplementary Concepts
  - Publication Type

1. Organize, Explore and Filter, using “Tags”

- **SWIFT tags documents according to:**
  - Health Outcomes

How does SWIFT assign Health Outcome tags?

- SWIFT contains “fingerprints” for each health-related MeSH term
- Fingerprints are generated by sampling citations from each of the top-level health-related MeSH terms in PubMed

Top 10 fingerprint terms for Mesh Code C09 “Otorhinolaryngologic Diseases”

Word	Type	Count	CorpusDoc Freq	tf_icdf_norm _alt
hear	Title	353	576	0.093727048
otitis media	Title 2-Gram	142	183	0.083971878
deaf	Title	211	328	0.075790822
otiti	Title	172	230	0.075037771
Deafness	MESH	387	637	0.074426499
ear	Title	197	311	0.071831078
laryng	Title	204	354	0.071168189
Laryngeal Neoplasms	MESH	388	664	0.070547647
glue ear	Title 2-Gram	9	9	0.070311944
nasal	Title	249	514	0.067404153

1. Use ‘Tags’ to organize, explore and filter

- **SWIFT tags documents according to:**
  - **Health Outcomes**
  - Evidence Stream
  - Exposure

  How does SWIFT assign Evidence Stream or Exposure tags?

  - Targeted queries for specific combinations of terms were developed to tag documents according to evidence stream and exposure. Queries were developed in collaboration with Stephanie Holmgren.

Example: Exposure Tag: Air Pollution

SWIFT Text Mining Workbench - [D:\Brians_Data\Documents\SWIFT Data\Global DNA Methylation_3.stp]

File Tools Preferences

Search | Browse MeSH Tree | Term Browser | Document Folders

Query:

```
( mesh_mh:( "air pollution" OR "air pollutants" OR "particulate matter" OR smog OR soot OR
 "vehicle emissions" OR motor vehicles )) OR
pharm_actions:"air pollutants" OR
( tiab: ( "air pollution" OR "air pollutant" OR "air pollutants" OR "particulate matter" O
R "PM2.5" OR "PM(2.5)" OR PM10 OR "PM(10)" OR smog OR soot OR "carbon black" OR "black car
bon" OR "elemental carbon" )) OR
( tiab:(( air OR airborne OR coarse OR ultrafine OR fine ) AND
          ( particle* OR particulate*))) OR
( tiab:(( vehicle OR vehicles OR vehicular OR auto OR automobile OR bus OR buses OR car OR
 truck* OR engine OR traffic OR transport* ) AND
        ( emissions OR exhaust OR fume* )))  OR

(( tiab:( air OR outdoor* OR outside OR ambient OR pollut* OR emissions OR exhaust* )) AND
```

Clear | Execute | Batch Query...

Document Preview | Fingerprint | Word Cloud | Pie Chart | Ranking Performance

Genomics and the respiratory effects of air pollution exposure.

Holloway JW, Savarimuthu Francis S, Fong KM, Yang IA. *Respirology (Carlton, Vic.) (2012)*

▼Abstract

Adverse health effects from air pollutants remain important, despite improvement in air quality in the past few decades. The exact mechanisms of lung injury from exposure to air pollutants are not yet fully understood. Studying the genome (e.g. single-nucleotide polymorphisms (SNP)), epigenome (e.g. methylation of genes), transcriptome (mRNA expression) and microRNAome (microRNA expression) has the potential to improve our understanding of the adverse effects of air pollutants. Genome-wide association studies of SNP have detected SNP associated with respiratory phenotypes; however, to date, only candidate gene studies of air pollution exposure have been performed. Changes in epigenetic processes, such DNA methylation that leads to gene silencing without altering the DNA sequence, occur with air pollutant exposure, especially global and gene-specific methylation changes. Respiratory cell line and animal models demonstrate distinct gene

Showing 367 of 35119 loaded documents (1 selected; 60 total included; 127 total training docs.)

Score	Training Item?	Included?	PMID	Title	Year	Authors	Journal
0.114			22404320	Genomics and the respiratory effects of air pollution exposure.	2012	Holloway JW, Savarimuthu Francis S, Fong KM, Yang IA	Respirology (Carlton, Vic.)
0.104			23052194	Comparison of genotoxicant-modified transcriptomic responses in conventi...	2012	Doktorova TY, Ellinger-Ziegelbauer H, Vinken M, Vanha...	Archives of toxicology
0.09			22055874	Genetic and epigenetic variations in inducible nitric oxide synthase promot...	2012	Salam MT, Byun HM, Lurmann F, Breton CV, Wang X, E...	The Journal of allergy and clinical immunology
0.053			22237295	Air pollution and markers of coagulation, inflammation, and endothelial fun...	2012	Bind MA, Baccarelli A, Zanobetti A, Tarantini L, Suh H, ...	Epidemiology (Cambridge, Mass.)
0.037			21660454	Association of secondhand smoke exposures with DNA methylation in blad...	2011	Wilhelm-Benartzi CS, Christensen BC, Koestler DC, And...	Cancer causes & control : CCC
0.034			23566092	Exposure to airborne particulate matter is associated with methylation pat...	2013	Sofer T, Baccarelli A, Cantone L, Coull B, Maity A, Lin X...	Epigenomics
0.032			22591701	Particulate matter, DNA methylation in nitric oxide synthase, and childhoo...	2012	Breton CV, Salam MT, Wang X, Byun HM, Siegmund KD...	Environmental health perspectives
0.031			21150337	LKB1/PEA3/ΔNp63 pathway regulates PTGS-2 (COX-2) transcription in lun...	2010	Ratovitski EA	Oxidative medicine and cellular longevity
0.03			19136372	Rapid DNA methylation changes after exposure to traffic particles.	2009	Baccarelli A, Wright RO, Bollati V, Tarantini L, Litonjua ...	American journal of respiratory and critical care medicine
0.03			18980546	Gene by environment interaction in asthma.	2009	London SJ, Romieu I	Annual review of public health
0.029			23476046	Blood hypomethylation of inflammatory genes mediates the effects of met...	2013	Tarantini L, Bonzini M, Tripodi A, Angelici L, Nordio F, C...	Occupational and environmental medicine
0.028			22441141	Smoking induces differential miRNA expression in human spermatozoa: a ...	2012	Marczylo EL, Amoako AA, Konje JC, Gant TW, Marczylo...	Epigenetics : official journal of the DNA Methylation So...
0.027			23855992	Evolutionary age of repetitive element subfamilies and sensitivity of DNA ...	2013	Byun HM, Motta V, Panni T, Bertazzi PA, Apostoli P, Ho...	Particle and fibre toxicology
0.027			21208146	Ambient air pollution exposure and damage to male gametes: human stud...	2011	Somers CM	Systems biology in reproductive medicine
0.025			23640490	Current genetics and epigenetics of smoking/tobacco-related cardiovascul...	2013	Breitling LP	Arteriosclerosis, thrombosis, and vascular biology
0.024			22989067	Genomic impact of cigarette smoke, with application to three smoking-rel...	2012	Talikka M, Sierro N, Ivanov NV, Chaudhary N, Peck MJ, ...	Critical reviews in toxicology

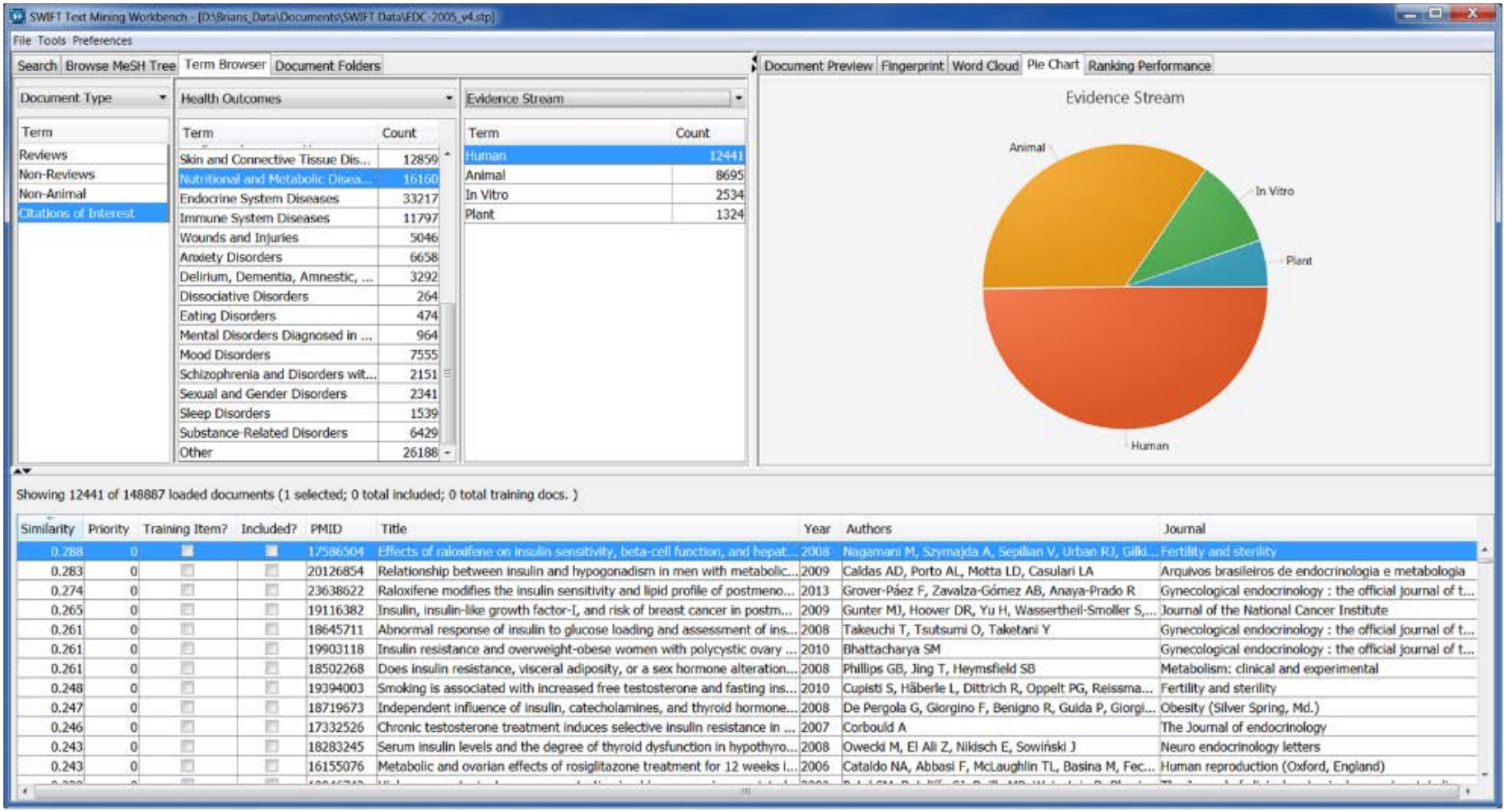
1. Use ‘Tags’ to organize, explore and filter

- **SWIFT tags documents according to:**
  - **Health Outcomes**
  - **Evidence Stream**
  - **Exposure**
  - Tox21 Chemicals

How does SWIFT assign Tox21 Chemical tags?

- Obtained list of 8,186 Tox21 chemicals from the EPA
- Also incorporated 2.7M synonyms for over 400,000 chemicals from public data
- On average, Tox21 chemicals had a mean of 20 synonyms

1. Use ‘Tags’ to organize, explore and filter



Chemical Classes by Health Outcome

Identify data rich/poor literature pockets at Chemical class vs. Health outcome intersection

	Bacterial Infections and Mycoses	Virus Diseases	Parasitic Diseases	Neoplasms	Musculoskeletal Diseases	Digestive System Diseases	Stomatognathic Diseases	Respiratory Tract Diseases	Otorhinolaryngologic Diseases	Nervous System Diseases	Eye Diseases	Male Urogenital Diseases	Female Urogenital Diseases and Preg	Cardiovascular Diseases	Hemic and Lymphatic Diseases	Congenital, Hereditary, and Neonatal	Skin and Connective Tissue Diseases	Nutritional and Metabolic Diseases	Endocrine System Diseases	Immune System Diseases	Wounds and Injuries	Anxiety Disorders	Delirium, Dementia, Amnestic, Cogni	Dissociative Disorders	Eating Disorders	Mental Disorders Diagnosed in Childh	Mood Disorders	Schizophrenia and Disorders with Psy	Sexual and Gender Disorders	Sleep Disorders	Substance-Related Disorders	Other
1. Indust-Intermediates	119	60	54	205	56	161	120	124	27	55	49	184	184	47	43	75	56	120	114	63	60	11	13	1	0	4	8	2	6	4	30	1136
1. Indust-Solvent	547	347	245	1311	927	733	311	700	271	1103	168	468	533	233	554	162	454	254	259	759	382	91	85	1	4	8	38	10	11	16	244	2857
1. Indust-Surfactant	37	16	7	65	26	144	7	12	9	61	4	100	89	11	21	36	71	62	254	38	5	12	4	0	0	2	3	0	18	0	3	394
1. Indust-Perchlorate	58	3	5	10	15	17	41	4	14	19	24	63	64	10	6	26	15	35	143	7	9	5	1	0	0	2	0	0	2	0	12	968
1. Indust-Octachlorostyre	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	6
2. Plastics-Phthalate	40	11	6	127	30	165	21	72	39	57	19	408	379	36	41	214	71	186	314	90	32	24	9	1	0	9	3	2	12	0	5	381
2. Plastics-Monomer	124	63	44	368	249	327	154	103	52	332	64	536	777	191	94	411	299	396	786	237	95	65	36	2	0	26	21	6	46	10	19	1102
2. Plastics-Dibutyltin	16	10	10	34	28	61	2	10	1	10	3	16	19	2	23	22	2	17	47	33	10	2	0	0	0	0	0	0	6	0	1	99
2. Plastics-Plasticizer	0	1	1	1	2	5	2	0	0	0	1	8	7	0	0	3	2	6	5	1	0	0	0	0	0	0	0	0	1	0	0	28
2. Plastics-Polymer/PVC	27	7	9	11	11	21	7	19	9	9	20	11	12	14	12	3	5	5	6	10	11	0	0	0	0	0	0	1	0	0	2	205
3. Flame Ret-Brominated	27	16	11	23	41	72	3	20	5	29	5	22	36	8	8	22	24	65	81	18	10	5	1	0	3	0	0	1	0	0	0	337
3. Flame Ret-PBDE	19	13	11	32	25	125	6	21	4	68	3	36	87	9	3	63	29	81	108	20	11	14	12	0	0	7	1	0	1	0	0	251
3. Flame Ret-BB 153	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
3. Flame Ret-Triphenyl phosphi	0	2	1	0	0	1	0	1	1	1	0	3	2	0	1	1	2	3	4	1	0	0	0	0	0	0	0	0	0	0	0	26
4. Metals-Arsenic, Cadmium, Le	572	230	257	1263	900	1788	383	646	179	1347	461	1994	2321	690	537	689	425	1465	691	557	453	441	274	1	52	130	28	17	14	33	388	7304
5. Cyclic Siloxanes-D4, D5, D6	3	0	0	2	1	8	0	11	4	1	0	1	5	0	0	1	4	6	4	0	1	0	1	0	0	0	0	0	0	0	0	68
6. PAH	196	182	51	2045	133	1169	257	572	54	112	154	295	443	138	261	154	905	436	361	349	94	103	46	0	7	8	6	2	2	2	111	1788
7. PFAS	19	40	6	90	61	302	21	45	16	87	7	122	227	53	38	149	51	185	129	75	28	16	18	0	1	15	0	0	4	0	20	610
8. Planar PCB	27	35	8	118	44	224	15	32	11	129	6	102	163	49	31	111	50	178	168	63	14	31	29	0	1	11	0	3	5	1	3	178
9 .Dioxins	62	93	41	468	121	523	72	135	28	130	18	193	360	127	123	235	214	326	246	230	59	35	21	0	9	13	3	4	8	3	16	1088
10. Pesticides-Fungicides	118	49	67	159	78	194	30	51	12	89	34	203	203	29	82	168	80	187	293	120	11	31	9	1	2	5	7	2	60	1	11	1150
10. Pesticides-Herbicides	169	56	65	140	56	170	21	54	28	130	38	209	218	67	72	97	50	131	240	103	33	54	16	1	5	4	4	2	14	0	15	1375
10. Pesticides-Microbiocides	433	97	113	62	26	79	181	26	14	28	19	72	68	29	6	22	50	36	89	42	39	5	5	0	0	0	0	0	0	1	4	206
10. Pesticides-Other	60	36	54	70	18	133	8	25	9	31	1	84	102	16	36	30	35	73	61	51	14	21	5	0	1	0	1	0	3	2	12	287
11. Steroids-Natural Hormone	1138	1139	731	7741	4386	3887	928	1101	559	5708	933	8662	10467	4809	1494	3983	4975	7985	23875	3681	1519	1438	1104	64	193	295	983	357	1735	536	781	1968
11. Steroids-Phytoestrogens	202	107	76	1095	387	599	50	147	36	270	74	488	540	339	212	178	614	690	662	266	141	63	85	3	1	4	16	10	16	6	24	330
11. Steroids-Synthetic Hormone	1442	1160	848	3881	2196	1825	519	1670	869	2660	1093	1236	2870	2290	2804	1359	1061	1915	3425	3730	1029	485	280	51	42	59	431	130	156	184	213	1283
12. Pharmaceuticals	725	642	384	5626	2440	2235	778	1464	675	6655	963	2112	2915	2879	1131	1216	4628	2039	3562	1611	1140	3854	1321	143	155	393	6173	1635	384	753	4622	1839
13. Consumer-UV Filters	3	1	2	8	5	9	0	0	3	8	3	8	9	0	0	6	28	16	12	4	5	1	0	0	0	0	0	0	0	0	1	51
13. Consumer-Food Additive	42	17	8	73	41	131	41	18	11	140	20	39	65	58	23	53	19	228	148	53	23	31	11	0	7	0	1	1	0	10	13	155

Grid | Raw | Health Categories

2. Identify over-represented concepts, terms, phrases

For a given corpus, visually investigate over-represented terms/phrases via **Word Cloud**

SWIFT Text Mining Workbench - [L:\Sciome\Projects\OHAT - SWIFT\NIEHS Datasets\IARC - DDT and Benzene\DDT_and_Benzene_2.stp]

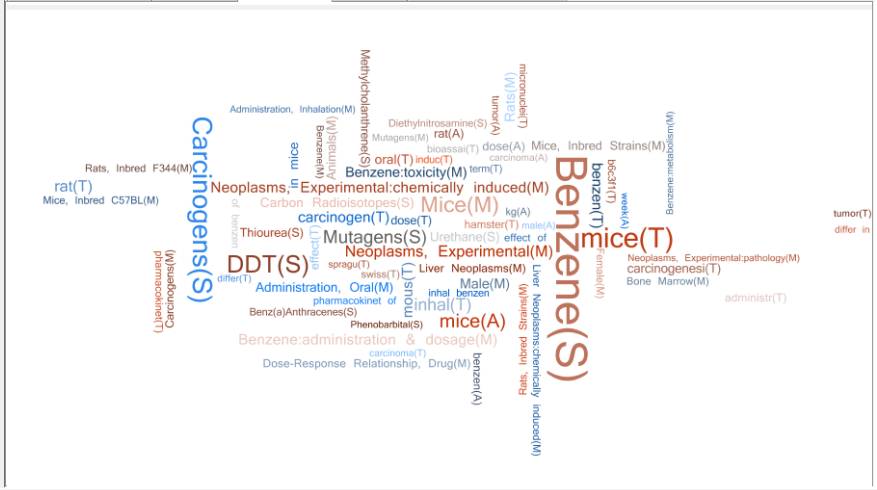
File Tools Help

Search | Browse MeSH Tree | Term Browser | Document Folders

Topic Models

Term	Code(s)	Count
Topic 46: study, results, data, present, high, analysis, found, time, poten...		13843
Topic 9: effects, environmental, health, chemical, chemicals, studies, hu...		3901
Topic 23: benzene, poisoning, chronic, blood, experimental, acute, patie...		2601
Topic 17: method, gas, determination, detection, benzene, extraction, a...		2196
Topic 22: ddt, pesticides, insecticides, chlorinated, residues, hydrocarbo...		2186
Topic 5: liver, rats, ddt, activity, effect, rat, hepatic, metabolism, enzym...		2126
Topic 16: ddt, milk, human, residues, samples, levels, organochlorine, pe...		1965
Topic 33: model, data, risk, assessment, models, exposure, based, para...		1763
Topic 32: ddt, dde, ddd, trichloro, metabolites, bis, p,p'-ddt, dichloro, bis...		1560
Topic 11: benzene, exposure, toluene, workers, ppm, concentrations, ex...		1477
Topic 2: benzene, removal, rate, concentration, treatment, water, btex, ...		1413
Topic 35: malaria, control, ddt, spraying, eradication, africa, health, vect...		1407
Topic 15: benzene, metabolites, metabolism, phenol, hydroquinone, met...		1362
Topic 4: exposure, benzene, risk, leukemia, cancer, study, workers, occu...		1359
Topic 13: ddt, effects, rats, effect, exposure, male, quail, adult, fed, day		1333
Topic 42: mice, benzene, rats, mg/kg, dose, mouse, male, doses, animal...		1293
Topic 45: benzene, exposure, urinary, urine, acid, workers, levels, expos...		1278

Document Preview | Fingerprint | Word Cloud | Pie Chart | Ranking Performance



Showing 1293 of 14443 loaded documents (1 selected; 0 total included; 0 total training docs.)

Similarity	Priority	Training Item?	Included?	PMID	Title	Year	Authors	Journal
0.123	0			25169096	[Preventive effects of garlic oil against the benzene-induced hematoto...	2014	Xu Z, Wang H, Chen Y, Mao G, Hu Y, Zeng T, Xie K	Zhonghua lao dong wei sheng zhi ye bing za zhi = Z...
0.071	0			24913574	Variability of cytometric parameters in various clusters of interstitial e...	2014	Bokov DA, Shevlyuk NN, Abdil'danova AM	Bulletin of experimental biology and medicine
0.045	0			24680972	PTEN methylation involved in benzene-induced hematotoxicity.	2014	Yang J, Zuo X, Bai W, Niu P, Tian L, Gao A	Experimental and molecular pathology
0.038	0			24355586	Dichlorodiphenyltrichloroethane exposure induces the growth of hepat...	2014	Jin XT, Song L, Zhao JY, Li ZY, Zhao MR, Liu WP	Toxicology letters
0.037	0			24345702	Immunotherapeutic potential of recombinant ESAT-6 protein in mouse...	2014	Mir SA, Verma I, Sharma S	Immunology letters
0.019	0			25190304	Mobile selected ion flow tube mass spectrometry (SIFT-MS) devices a...	2014	Storer M, Salmond J, Dirks KN, Kingham S, Epton M	Journal of breath research
0.015	0			24820114	The evaluation of p,p'-DDT exposure on cell adhesion of hepatocellular...	2014	Jin X, Chen M, Song L, Li H, Li Z	Toxicology
0.008	0			24701929	Rendering plant emissions of volatile organic compounds during sterili...	2014	Bhatti ZA, Maqbool F, Langenhove HV	Environmental technology
0.007	0			24029690	Risk assessment of manufacturing equipment surfaces contaminated ...	2014	Luo F, Song J, Chen MF, Wei J, Pan YY, Yu HB	The Science of the total environment
0.004	0			24570016	Surgical smoke may be a biohazard to surgeons performing laparoscop...	2014	Choi SH, Kwon TG, Chung SK, Kim TH	Surgical endoscopy
0.002	0			24517295	Chamber studies on nonvented decorative fireplaces using liquid or gel...	2014	Schripp T, Salthammer T, Wientzek S, Wensing M	Environmental science & technology
0.002	0			25080070	Replacing fish meal by food waste in feed pellets to culture lower trop...	2014	Cheng Z, Mo WY, Man YB, Nie XP, Li KB, Wong MH	Environment international
0.001	0			25113182	Residues and chiral signatures of organochlorine pesticides in mollusk...	2014	Zhou S, Tang Q, Jin M, Liu W, Niu L, Ye H	Chemosphere
0.098	0			23433151	[The effects of benzene poisoning on expression of multidrug resistanc...	2013	Huang JS, Shi JM, Zhang JH, Li B, Fan W, Zhou YL	Zhonghua lao dong wei sheng zhi ye bing za zhi = Z...

2. Identify over-represented concepts, terms, phrases

For a given corpus, visually investigate over-represented terms/phrases via **Fingerprints**

SWIFT Text Mining Workbench - [D:\Brians_Data\Documents\SWIFT Data\Global DNA Methylation_3.stp]

File Tools Preferences

Search | Browse MeSH Tree | Term Browser | Document Folders

Health Outcomes | Edit...

- All Documents
  - Bacterial Infections and Mycoses
  - Virus Diseases
  - Parasitic Diseases
  - Neoplasms
  - Musculoskeletal Diseases
  - Digestive System Diseases
  - Stomatognathic Diseases
  - Respiratory Tract Diseases
  - Otorhinolaryngologic Diseases
  - Nervous System Diseases
  - Eye Diseases
  - Male Urogenital Diseases
  - Female Urogenital Diseases and Pregnancy Complications
  - Cardiovascular Diseases
  - Hemic and Lymphatic Diseases

Redistribute | Prioritize

Document Preview | Fingerprint | Word Cloud | Pie Chart | Ranking Performance

Ignore?	Word	Type	Word Count	Document Frequency	Score
	Escherichia coli	MESH	384	384	22.301
	coli	Title	136	136	16.07
	escherichia	Title	107	107	14.677
	coli	Abstract	828	559	13.036
	escherichia coli	Title 2-Gram	107	107	12.805
	bacteri	Title	76	75	11.357
	bacteri	Abstract	779	571	10.687
	pathogen	Title	84	81	8.721
	pathogen	Abstract	1074	769	8.535
	bacteria	Abstract	476	354	8.06
	pylori	Title	67	67	7.307
	helicobact	Title	64	64	7.192
	e coli	Abstract 2-...	334	203	7.118
	infect	Abstract	3014	1421	6.894
	pylori	Abstract	430	109	6.721
	candida	Title	26	26	6.343
	infect	Title	258	255	5.902

Showing 3254 of 35119 loaded documents (1 selected; 60 total included; 127 total training docs.)

Fingerprint uses 733 words.

Similarity	Priority	Training Item?	Included?	PMID	Title	Year	Authors	Journal
0.479	0			20122996	Helicobacter pylori infection generates genetic instability in gastric cells.	2010	Machado AM, Figueiredo C, Seruca R, Rasmussen LJ	Biochimica et biophysica acta
0.375	0			20345486	Role of Helicobacter pylori infection in aberrant DNA methylation along...	2010	Shin CM, Kim N, Jung Y, Park JH, Kang GH, Kim JS, J...	Cancer science
0.373	0			21631269	Aberrant CpG island hypermethylation in pediatric gastric mucosa in a...	2011	Shin SH, Park SY, Ko JS, Kim N, Kang GH	Archives of pathology & laboratory medicine
0.367	0			22937582	[Pathogenetic importance of helicobacter pylori infection].	2012	Sklianskaia OA, Lapina TL	Arkhiv patologii
0.358	0			20044995	CpG methylation and reduced expression of O6-methylguanine DNA m...	2010	Sepulveda AR, Yao Y, Yan W, Park DI, Kim JJ, Goodi...	Gastroenterology
0.337	0			20602342	Alu and Sato hypomethylation in Helicobacter pylori-infected gastric m...	2011	Yoshida T, Yamashita S, Takamura-Enya T, Niwa T, ...	International journal of cancer. Journal international...
0.335	0			21585603	Genome-wide DNA methylation profiles in noncancerous gastric muco...	2011	Shin CM, Kim N, Jung Y, Park JH, Kang GH, Park WY,...	Helicobacter
0.333	0			19639607	Comparison of CpG island hypermethylation and repetitive DNA hypom...	2009	Park SY, Yoo EJ, Cho NY, Kim N, Kang GH	The Journal of pathology
0.326	0			19035455	The presence of a methylation fingerprint of Helicobacter pylori infecti...	2009	Nakajima T, Yamashita S, Maekita T, Niwa T, Nakaz...	International journal of cancer. Journal international...
0.315	0			18755836	Identification of pseudouridine methyltransferase in Escherichia coli.	2008	Ero R, Peil L, Liiv A, Remme J	RNA (New York, N.Y.)
0.311	0			22211478	Genome-wide transcriptome analysis of fluoroquinolone resistance in c...	2012	Yamane T, Enokida H, Hayami H, Kawahara M, Naka...	International journal of urology : official journal of th...
0.31	0			6986915	Hybridization analysis of the methylated bases of Escherichia coli DNA.	1980	Quint A, Cedar H	Biochimica et biophysica acta
0.306	0			16940095	Evidence of antibiotic resistance gene silencing in Escherichia coli.	2006	Enne VI, Delsol AA, Roe JM, Bennett PM	Antimicrobial agents and chemotherapy
0.3	0			23346103	Association between Genetic Instability and Helicobacter pylori Infecti...	2012	Kim JS, Chung WC, Lee KM, Paik CN, Lee KS, Kim HJ...	Gastroenterology research and practice
0.295	0			23130834	Epigenetic implication of gene-adjacent retroelements in Helicobacter ...	2012	Rhyu MG, Oh JH, Hong SJ	Epigenomics
0.293	0			6319759	Evidence that Escherichia coli virus T1 induces a DNA methyltransfera...	1984	Auer B, Schweiger M	Journal of virology

3. Search using advanced query syntax

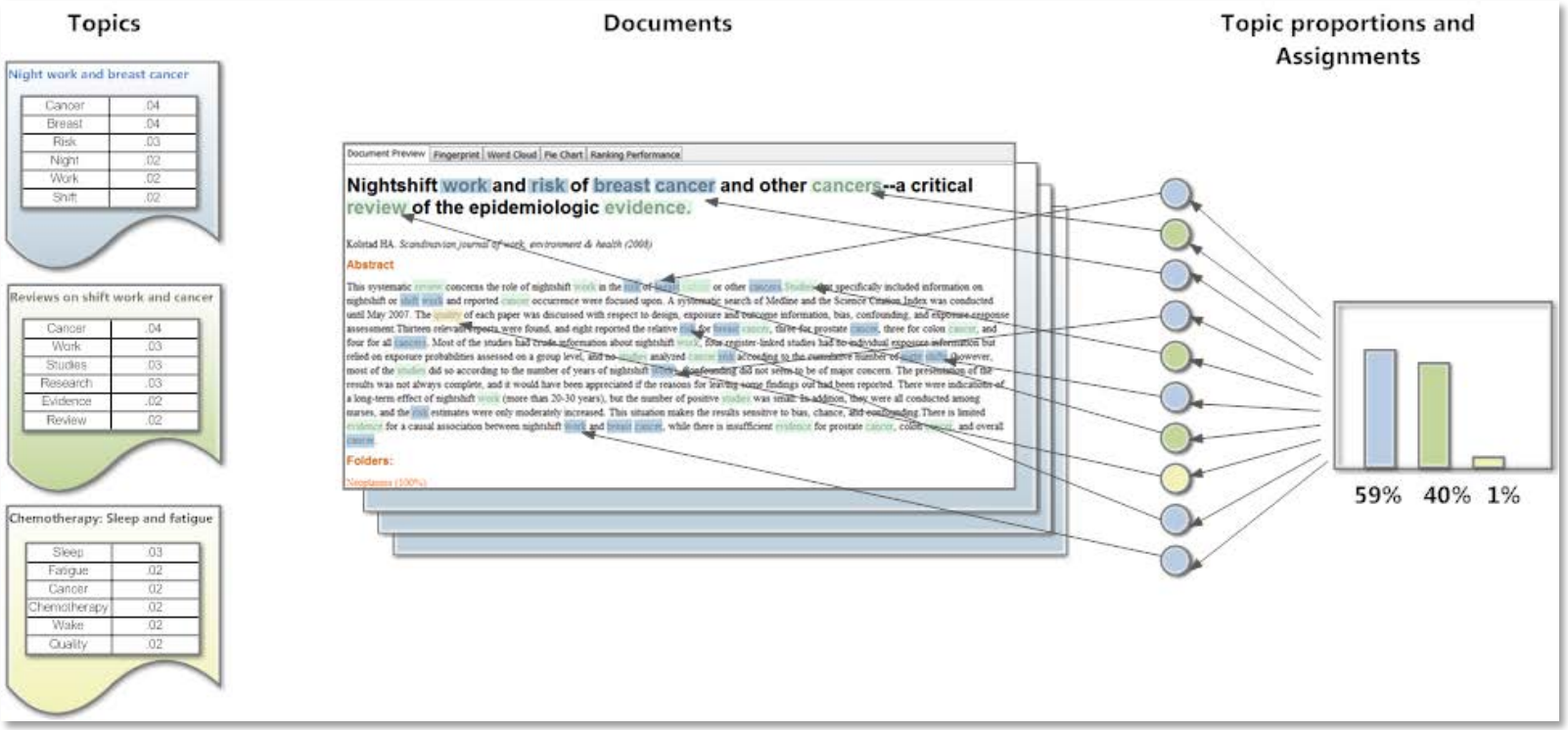
- SWIFT employs the powerful *Lucene* data indexing engine which uses an “inverted index” to map document sections to words for extremely fast lookups

- Advanced querying capability such as:
  - Boolean operators — human AND cancer
  - Wildcards — optimiz* AND nucleo?ide
  - Proximity search — "first generation"~5
  - Fuzzy match — exposure~
  - Fielded searching — title: “estrogen receptor”
  - Ranged queries — pubyear:[2010 TO 2014]

- **Full text searching capabilities (full text vs. Abstract/Titles)**

4. Categorize according to relevant sub-topics

Topic Modeling



Exploratory tool, useful for understanding the topics contained in large document set when no prior knowledge is available

4. Categorize according to relevant sub-topics

Topic Modeling

- We use the Latent Dirichlet Allocation (LDA) topic modeling approach (Blei 2003) to probabilistically assign documents to topics.
- Under this framework, "topics" are conceptualized as probability distributions over a vocabulary; documents are "bags of words" randomly generated conditional on these hidden topics.
- LDA provides a statistical framework which can be used to cluster related documents into meaningful topics.
- It is an unsupervised algorithm: topic membership and also the topics themselves are discovered "automatically" from an unlabeled corpus

Reference: David M. Blei, Andrew Y. Ng and Michael I. Jordan; Journal of Machine Learning Research 3 (2003) 993-1022.

Topic Modeling Example: Leukemia and Benzene exposure

SWIFT Text Mining Workbench - [L:\Sciome\Projects\OHAT - SWIFT\NIEHS Datasets\IARC - DDT and Benzene\DDT_and_Benzene_2.stp]

File Tools Help

Search | Browse MeSH Tree | Term Browser | Document Folders

Topic Models

Term	Count
Topic 46: study, results, data, present, high, analysis, found, time, potential, observed	13843
Topic 9: effects, environmental, health, chemical, chemicals, studies, human, toxic, risk, exposure	3901
Topic 23: benzene, poisoning, chronic, blood, experimental, acute, patients, benzol, case, solvents	2601
Topic 17: method, gas, determination, detection, benzene, extraction, analysis, samples, chromatog...	2196
Topic 22: ddt, pesticides, insecticides, chlorinated, residues, hydrocarbon, dieldrin, organochlorine, p...	2186
Topic 5: liver, rats, ddt, activity, effect, rat, hepatic, metabolism, enzymes, microsomal	2126
Topic 16: ddt, milk, human, residues, samples, levels, organochlorine, pesticide, tissue, pesticides	1965
Topic 33: model, data, risk, assessment, models, exposure, based, parameters, approach, benzene	1763
Topic 32: ddt, dde, ddd, trichloro, metabolites, bis, p,p'-ddt, dichloro, bis(p-chlorophenyl)ethane, deg...	1560
Topic 11: benzene, exposure, toluene, workers, ppm, concentrations, exposures, occupational, solve...	1477
Topic 2: benzene, removal, rate, concentration, treatment, water, btex, organic, sorption, phase	1413
Topic 35: malaria, control, ddt, spraying, eradication, africa, health, vector, areas, irs	1407
Topic 15: benzene, metabolites, metabolism, phenol, hydroquinone, metabolite, liver, formation, cat...	1362
Topic 4: exposure, benzene, risk, leukemia, cancer, study, workers, occupational, studies, cases	1359
Topic 13: ddt, effects, rats, effect, exposure, male, quail, adult, fed, day	1333
Topic 42: mice, benzene, rats, mg/kg, dose, mouse, male, doses, animals, inhalation	1293
Topic 45: benzene, exposure, urinary, urine, acid, workers, levels, exposed, blood, biological	1278

Document Preview | Fingerprint | Word Cloud | Pie Chart | Ranking Performance

Leukemia risk associated with low-level benzene exposure.

Glass DC, Gray CN, Jolley DJ, Gibbons C, Sim MR, Fritschi L, Adams GG, Bisby JA, Manuell R.
Epidemiology (Cambridge, Mass.) (2003)

▼Abstract

Men who were part of an Australian petroleum industry cohort had previously been found to have an excess of lympho-hematopoietic cancer. Occupational benzene exposure is a possible cause of this excess.We conducted a case-control study of lympho-hematopoietic cancer nested within the existing cohort study to examine the role of benzene exposure. Cases identified between 1981 and 1999 (N = 79) were age-matched to 5 control subjects from the cohort. We estimated each subject's benzene exposure using occupational histories, local site-specific information, and an algorithm using Australian petroleum industry monitoring data.Matched analyses showed that the risk of leukemia was increased at cumulative exposures above 2 ppm-years and with intensity of exposure of highest exposed job over 0.8 ppm. Risk increased with higher exposures; for the 13 case-sets with greater than 8 ppm-years cumulative exposure, the odds ratio was 11.3 (95% confidence interval = 2.85-45.1). The risk of

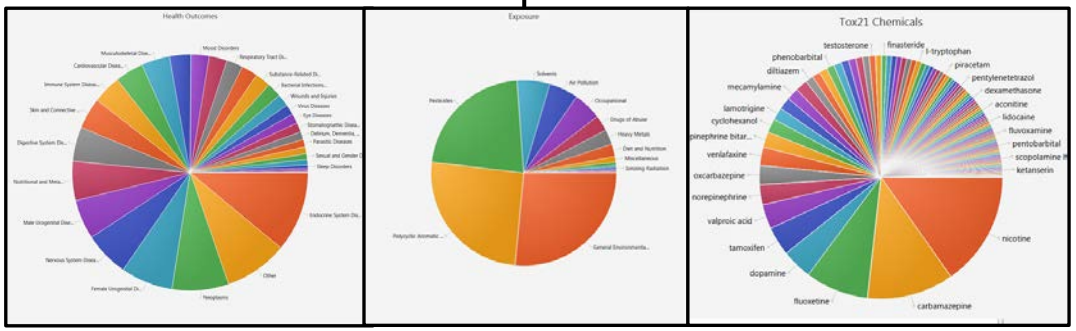
Showing 1359 of 14443 loaded documents (1 selected; 0 total included; 0 total training docs.)

Similarity	Priority	Training Item?	Included?	PMID	Title	Year	Authors	Journal
0.292	0			14501272	Leukemia risk associated with low-level benzene exposure.	2003	Glass DC, Gray CN, Jolley DJ, Gibbons C, Sim MR, Fr...	Epidemiology (Cambridge, Mass.)
0.271	0			7345926	Leukemia in benzene workers.	1981	Rinsky RA, Young RJ, Smith AB	American journal of industrial medicine
0.268	0			3890530	Projections of leukemia risk associated with occupational exposure to ...	1985	Infante PF, White MC	American journal of industrial medicine
0.263	0			12125084	Leukemia in relation to occupational exposures to benzene and other a...	2002	Guénel P, Imbernon E, Chevalier A, Crinquand-Calas...	American journal of industrial medicine
0.263	0			2792040	Benzene and leukemia: an epidemiologic risk assessment.	1989	Rinsky RA	Environmental health perspectives
0.252	0			3561457	Benzene and leukemia. An epidemiologic risk assessment.	1987	Rinsky RA, Smith AB, Hornung R, Filloon TG, Young ...	The New England journal of medicine
0.245	0			2611155	Risk assessment of leukaemia and occupational exposure to benzene.	1989	Swaen GM, Meijers JM	British journal of industrial medicine
0.238	0			9118924	Leukemia mortality by cell type in petroleum workers with potential ex...	1996	Raabe GK, Wong O	Environmental health perspectives
0.237	0			2792042	A retrospective cohort study of leukemia and other cancers in benzene...	1989	Yin SN, Li GL, Tain FD, Fu ZI, Jin C, Chen YJ, Luo SJ,...	Environmental health perspectives
0.236	0			9118929	Leukemia risk associated with benzene exposure in the Pliofilm cohort.	1996	Paxton MB	Environmental health perspectives
0.228	0			8008923	Leukemia risk associated with benzene exposure in the pliofilm cohort:...	1994	Paxton MB, Chinchilli VM, Brett SM, Rodricks JV	Risk analysis : an official publication of the Society f...
0.227	0			9155776	A case-control study to investigate the risk of leukaemia associated wi...	1997	Rushton L, Romaniuk H	Occupational and environmental medicine
0.224	0			8008924	Leukemia risk associated with benzene exposure in the pliofilm cohort....	1994	Paxton MB, Chinchilli VM, Brett SM, Rodricks JV	Risk analysis : an official publication of the Society f...
0.222	0			9115030	Cancer mortality among workers with benzene exposure.	1997	Ireland B, Collins JJ, Buckley CF, Riordan SG	Epidemiology (Cambridge, Mass.)

Outline

[A] What topics are being discussed in this corpus?

1. Organize and explore documents via tagging
2. Identify over-represented concepts, terms, phrases
3. Query using advanced syntax
4. Categorize according to relevant sub-topics



Scoping Report

Problem Formulation

Topic Refinement

Outline

[A] What topics are being discussed in this corpus?

1. Organize and explore documents via tagging
2. Identify over-represented concepts, terms, phrases
3. Query using advanced syntax
4. Categorize according to relevant sub-topics

[B] How to facilitate manual screening?

5. Use machine learning to rank documents
6. Perform iterative ranking to increase recall

7. Future directions

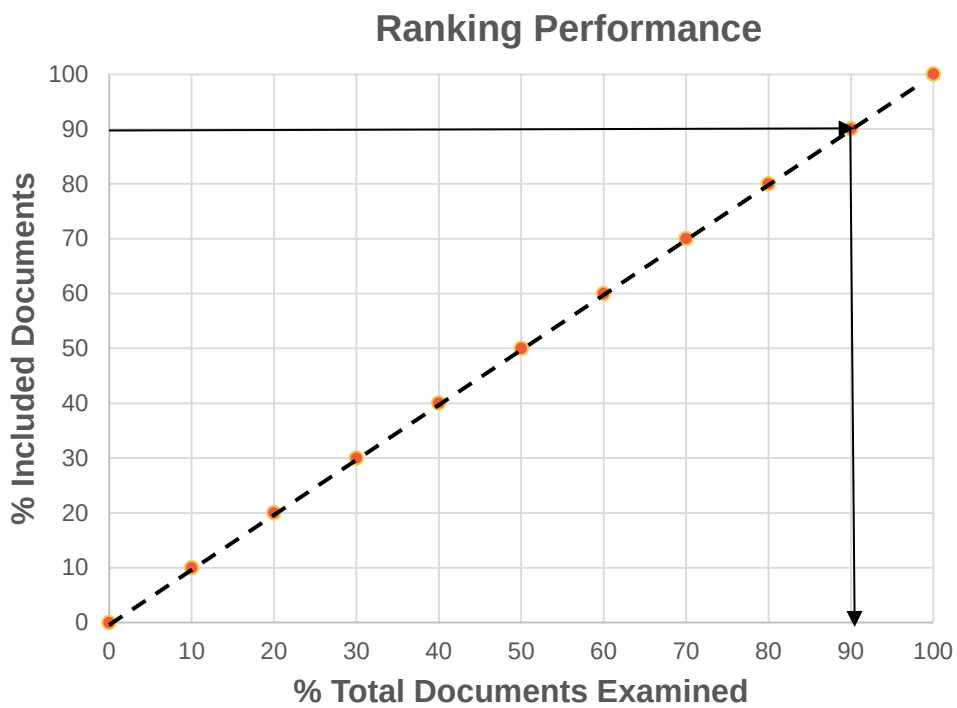
5. Use Machine Learning to Rank Documents

Use Seed Documents → Generate Distinguishing Features & Rank → Screen

- SWIFT uses machine learning and user provided small "seed set" of *included* or *excluded* documents to build models
- The idea is to 'learn' distinguishing features between *included* and *excluded* seed set documents and use it to rank order rest of the corpus.
- SWIFT uses a variety of document features to perform the modeling
  - Bag of words, bi/tri-grams, annotations etc.
  - Topic-model weights
- Documents are ranked probabilistically using a regularized, log-linear model (very similar in practice to Support Vector Machine with a linear kernel)
- Results are benchmarked using data from previous reviews:
  - Computer generated ranking is compared to inclusion ascertained by reviewer

5. Use Machine Learning to Rank Documents

If screening is performed without ordering the documents in any specific manner, %screened and %Recall would approximately follow the 45 degree dotted black line (----)



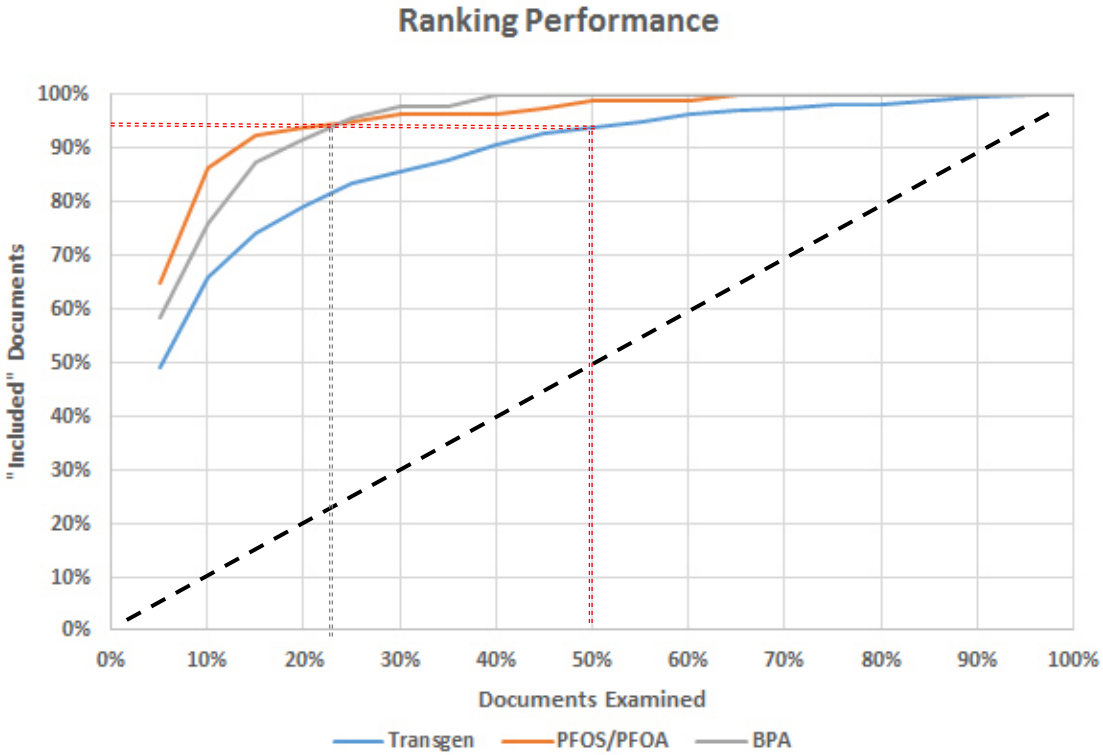
X-axis: Percent of total documents examined
Y-axis: Percent of *Included* documents discovered
For this experiment, seed = 30 included + 30 excluded were used

		PFOS/PFOA	BPA	Transgenerational
PubMed	**Total**	6,331	7,700	48,638
	Included	95 (1.5%)	111 (1.4%)	765 (1.6%)

5. Use Machine Learning to Rank Documents

Conclusions:

- The top of the resulting ranked lists contain the majority of the relevant documents.
- >95% Recall can be achieved with approx. 50% or less screening.

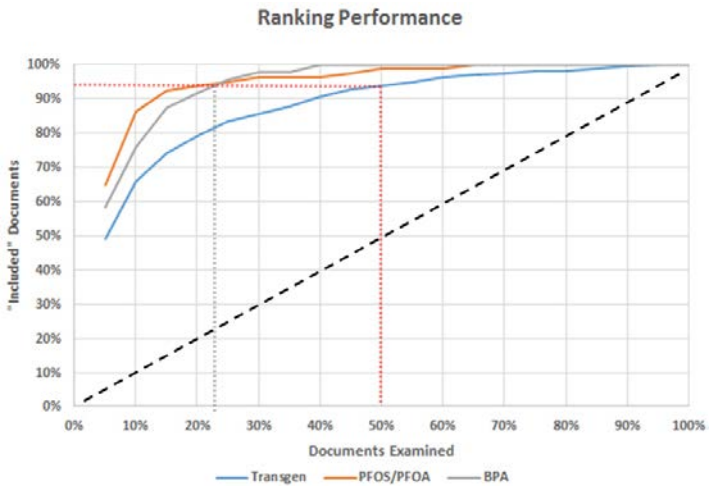


X-axis: Percent of total documents examined
Y-axis: Percent of *Included* documents discovered
For this experiment, seed = 30 included + 30 excluded were used

		PFOS/PFOA	BPA	Transgenerational
PubMed	**Total**	6,331	7,700	48,638
	Included	95 (1.5%)	111 (1.4%)	765 (1.6%)

6. Iterative Ranking to Increase Recall

A key consideration in use of machine learning for reducing screening burden is to ensure a very high recall.



How to know when to stop screening?
How do you ensure high Recall?
We are actively developing and deploying methods to address these questions:

i. First approach: A combination of Ranking and Random sampling combined with basic statistics (binomial or geometric distribution) to make probabilistic argument for stopping.
ii. Build model that explicitly models trade-off between precision and recall with a user–input parameter

Improve and enrich seed set via **iteratively ranking** and screening:

- incorporate topic modeling, clustering and or random sampling when choosing seeds

Under Development

In Summary

SWIFT provides a user-friendly workbench to:

Explore, Categorize, Search, Topic discovery→ Scoping Report

Prioritize Documents to facilitates screening → Efficient Screening

Work under progress:

- Further develop document prioritization modules
- Use of FIDDLE for Full-Text extraction
- Integration with HAWC

Acknowledgement

A Team effort led by:
Brian Howard
Jason Phillips
Ruchir Shah

With many valuable contributions from:
Kyle Miller
Deepak Mav
Mihir Shah

And an amazing ongoing collaboration and supervision:
Vickie Walker, Katie Pelch, Abee Boyles
Stephanie Holmgren, Andrew Rooney, Kristina Thayer